What Is Claimed Is:

1   1.   A method for selecting a node to host a primary server for a service
2   from a plurality of nodes in a distributed computing system, the method
3   comprising:
4   receiving an indication that a state of the distributed computing system has
5   changed;
6   in response to the indication, determining if there is already a node hosting
7   the primary server for the service; and
8   if there is not already a node hosting the primary server, selecting a node to
9   host the primary server based upon rank information for the nodes.

1   2.   The method of claim 1, wherein selecting the node to host the
2   primary server involves:
3   assuming that a given node from the plurality of nodes in the distributed
4   computing system hosts the primary server,
5   communicating rank information between the given node and other nodes
6   in the distributed computing system, wherein each node in the distributed
7   computing system has a unique rank with respect to the other nodes in the
8   distributed computing system,
9   comparing a rank of the given node with a rank of the other nodes in the
10   distributed computing system, and
11   if one of the other nodes in the distributed computing system has a higher
12   rank than the given node, disqualifying the given node from hosting the primary
13   server.

1   3.   The method of claim 2, further comprising, if there exists a node

2   that is configured to host the primary server, allowing the node that is configured

3   to host the primary server to communicate with other nodes in the distributed

4   computing system in order to disqualify the other nodes from hosting the primary

5   server.

1   4.   The method of claim 2, wherein assuming that the given node

2   hosts the primary server involves:

3   maintaining a candidate variable in the given node identifying a candidate

4   node to host the primary server; and

5   initially setting the candidate variable to identify the given node.

1   5.   The method of claim 1, further comprising, after a new node has

2   been selected to host the primary server, if the new node is different from a

3   previous node that hosted the primary server, establishing connections for the

4   service to the new node.

1   6.   The method of claim 1, further comprising, after a new node has

2   been selected to host the primary server, if the new node is different from a

3   previous node that hosted the primary server, configuring the new node to host the

4   primary server for the service.

1   7.   The method of claim 1, further comprising restarting the service if

2   the service was interrupted as a result of the change in state of the distributed

3   computing system.

14

1   8.   The method of claim 2, wherein the given node in the distributed
2 computing system acts a one of:
3   a host for the primary server for the service;
4   a host for a secondary server for the service, wherein the secondary server
5 periodically receives checkpointing information from the primary server; and
6   a spare for the primary server, wherein the spare does not receive
7 checkpointing information from the primary server.

1   9.   The method of claim 8, further comprising, upon initial startup of
2 the service, selecting a highest ranking spare to host the primary server for the
3 service.

1   10.   The method of claim 8, further comprising allowing the primary
2 server to configure spares in the distributed computing system to host secondary
3 servers for the service.

1   11.   The method of claim 8, wherein comparing the rank of the given
2 node with the rank of the other nodes in the distributed computing system
3 involves considering a host for the primary server to have a higher rank than a
4 host for a space, and considering a host for a secondary server to have a higher
5 rank than a spare.

1   12.   The method of claim 2, wherein disqualifying the given node from
2 hosting the primary server involves ceasing to communicate rank information
3 between the given node and the other nodes in the distributed computing system.

15

1   13.   A computer-readable storage medium storing instructions that

2   when executed by a computer cause the computer to perform a method for

3   selecting a node to host a primary server for a service from a plurality of nodes in

4   a distributed computing system, the method comprising:

5   receiving an indication that a state of the distributed computing system has

6   changed;

7   in response to the indication, determining if there is already a node hosting

8   the primary server for the service; and

9   if there is not already a node hosting the primary server, selecting a node to

10   host the primary server based upon rank information for the nodes.

1   14.   The computer-readable storage medium of claim 13, wherein

2   selecting the node to host the primary server involves:

3   assuming that a given node from the plurality of nodes in the distributed

4   computing system hosts the primary server,

5   communicating rank information between the given node and other nodes

6   in the distributed computing system, wherein each node in the distributed

7   computing system has a unique rank with respect to the other nodes in the

8   distributed computing system,

9   comparing a rank of the given node with a rank of the other nodes in the

10   distributed computing system, and

11   if one of the other nodes in the distributed computing system has a higher

12   rank than the given node, disqualifying the given node from hosting the primary

13   server.

1   15.   The computer-readable storage medium of claim 14, wherein if

2   there exists a node that is configured to host the primary server, the method

3   further comprises allowing the node that is configured to host the primary server

4   to communicate with other nodes in the distributed computing system in order to

5   disqualify the other nodes from hosting the primary server.

1   16.   The computer-readable storage medium of claim 14, wherein

2   assuming that the given node hosts the primary server involves:

3   maintaining a candidate variable in the given node identifying a candidate

4   node to host the primary server; and

5   initially setting the candidate variable to identify the given node.

1   17.   The computer-readable storage medium of claim 13, wherein after

2   a new node has been selected to host the primary server, if the new node is

3   different from a previous node that hosted the primary server, the method further

4   comprises establishing connections for the service to the new node.

1   18.   The computer-readable storage medium of claim 13, wherein after

2   a new node has been selected to host the primary server, if the new node is

3   different from a previous node that hosted the primary server, the method further

4   comprises configuring the new node to host the primary server for the service.

1   19.   The computer-readable storage medium of claim 13, wherein the

2   method further comprises restarting the service if the service was interrupted as a

3   result of the change in state of the distributed computing system.

1   20.   The computer-readable storage medium of claim 14, wherein the

2   given node in the distributed computing system acts a one of:

3   a host for the primary server for the service;

17

1   a host for a secondary server for the service, wherein the secondary server

2   periodically receives checkpointing information from the primary server; and

3   a spare for the primary server, wherein the spare does not receive

4   checkpointing information from the primary server.

1   21.   The computer-readable storage medium of claim 20, wherein upon

2   initial startup of the service, the method further comprises selecting a highest

3   ranking spare to host the primary server for the service.

1   22.   The computer-readable storage medium of claim 20, wherein the

2   method further comprises allowing the primary server to configure spares in the

3   distributed computing system to host secondary servers for the service.

1   23.   The computer-readable storage medium of claim 20, wherein

2   comparing the rank of the given node with the rank of the other nodes in the

3   distributed computing system involves considering a host for the primary server to

4   have a higher rank than a host for a space, and considering a host for a secondary

5   server to have a higher rank than a spare.

1   24.   The computer-readable storage medium of claim 14, wherein

2   disqualifying the given node from hosting the primary server involves ceasing to

3   communicate rank information between the given node and the other nodes in the

4   distributed computing system.

1   25.   An apparatus that selects a node to host a primary server for a

2   service from a plurality of nodes in a distributed computing system, the apparatus

3   comprising:

18

4   a receiving mechanism that is configured to receive an indication that a

5 state of the distributed computing system has changed;

6   a determination mechanism that is configured to determine if there is

7 already a node hosting the primary server for the service in response to the

8 indication;

9   a selecting mechanism, wherein if there is not already a node hosting the

10 primary server, the selecting mechanism is configured to select a node to host the

11 primary server based upon rank information for the nodes.

1   26.   The apparatus of claim 25, wherein, in selecting a node to host the

2 primary server based upon rank information, the selecting mechanism is

3 configured to:

4   communicate rank information between the given node and other nodes in

5 the distributed computing system, wherein each node in the distributed computing

6 system has a unique rank with respect to the other nodes in the distributed

7 computing system, and to

8   compare a rank of the given node with a rank of the other nodes in the

9 distributed computing system.

1   27.   The apparatus of claim 26, further comprising a disqualification

2 mechanism that is configured to disqualify the given node from hosting the

3 primary server if one of the other nodes in the distributed computing system has a

4 higher rank than the given node.

1   28.   The apparatus of claim 26, further comprising a mechanism on the

2 primary server that is configured to communicate with other nodes in the

3   distributed computing system in order to disqualify the other nodes from hosting

4   the primary server.

1   29.   The apparatus of claim 26, wherein the selecting mechanism is

2   configured to:

3   maintain a candidate variable in the given node identifying a candidate

4   node to host the primary server; and to

5   initially set the candidate variable to identify the given node.

1   30.   The apparatus of claim 25, further comprising a connection

2   mechanism that is configured to establish connections for the service to a new

3   node after the new node has been selected to host the primary server, and if the

4   new node is different from a previous node that hosted the primary server.

1   31.   The apparatus of claim 25, further comprising a mechanism that

2   configures a new node to host the primary server for the service, after the new

3   node has been selected to host the primary server, and if the new node is different

4   from a previous node that hosted the primary server.

1   32.   The apparatus of claim 25, further comprising a restarting

2   mechanism that is configured to restart the service if the service was interrupted as

3   a result of the change in state of the distributed computing system.

1   33.   The apparatus of claim 26, wherein the given node in the

2   distributed computing system acts a one of:

3   a host for the primary server for the service;

20

1   a host for a secondary server for the service, wherein the secondary server

2   periodically receives checkpointing information from the primary server; and

3   a spare for the primary server, wherein the spare does not receive

4   checkpointing information from the primary server.

1   34.   The apparatus of claim 33, further comprising an initialization

2   mechanism wherein during initialization of the service, the initialization

3   mechanism is configured to select a highest ranking spare to host the primary

4   server for the service.

1   35.   The apparatus of claim 33, further comprising a promotion

2   mechanism on the primary server that that is configured to promote spares in the

3   distributed computing system to host secondary servers for the service.

1   36.   The apparatus of claim 33, wherein while comparing the rank of

2   the given node with the rank of the other nodes in the distributed computing

3   system, the selecting mechanism is configured to consider a host for the primary

4   server to have a higher rank than a host for a secondary server, and to consider a

5   host for a secondary server to have a higher rank than a spare.

1   37.   The apparatus of claim 26, wherein the selecting mechanism is

2   configured to cease to communicate rank information between the given node and

3   the other nodes in the distributed computing system after the given node is

4   disqualified by the disqualification mechanism.

1   38.   A method for selecting a node to host a primary server for a service

2   from a plurality of nodes in a distributed computer system, comprising:

21

3   communicating disqualification information between the node and

4 remaining nodes in the plurality of nodes;

5   disqualifying the node from hosting the primary server based upon the

6 disqualification information received from the remaining nodes.

1   39.   The method of claim 38, wherein the disqualification information

2 comprises a node rank information.

1   40.   The method of claim 39, wherein the node rank for a given node is

2 calculated using an assumption that the given node hosts the primary server.

1   41.   The method of claim 40, wherein the calculated node rank is

2 unique with respect to the ranks of other nodes in the distributed computer system.

1   42.   The method of claim 39, wherein the disqualifying of the node

2 comprises:

3   comparing a rank of the node to a set of ranks of the remaining nodes in

4 the distributed computer system; and

5   disqualifying the node from hosting the primary server if one of the set of

6 ranks of the remaining nodes is higher than the rank of the node.

1   43.   The method of claim 38, further comprising repeating the acts of

2 communicating disqualification information and disqualifying the node for at least

3 one more node in the plurality of nodes.

22